DOCUMENT RESUME

ED 201 667 TM 810 261

AUTHOR Willson, Victor L.
TITLE Robinson's Measure of Agreement as a Parallel Forms Reliability Coefficient.
PUB DATE [77]
NOTE 11p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Comparative Analysis; *Correlation; *Difficulty Level; *Mathematical Formulas; Simulation; Test Items; *Test Reliability
IDENTIFIERS *Parallel Forms Reliability; *Robinsons Measure of Agreement

ABSTRACT

A major deficiency in classical test theory is the reliance on Pearson product-moment (PPM) correlation concepts in the definition of reliability. PPM measures are totally insensitive to first moment differences in tests which leads to the dubious assumption of essential tau-equivalence. Robinson proposed a measure of agreement that is sensitive to different test difficulty and gives a practical statistic to estimate reliability in the presence of known form variation in difficulty. Robinson's measure of agreement appears to be a useful alternative to the generalizability coefficient, as it provides a more conservative estimate of reliability under conditions of parallel form differences in mean. This is likely to be especially useful when examining inter rater reliability when internal consistency of the raters is poor. Robinson's measure does not seem advantageous for highly reliable parallel tests such as are encountered in standardized testing programs. A simulation study is presented to illustrate the degree of the coefficient's sensitivity to form difficulty variance. Robinson's measure of agreement and the intraclass correlation are computed for each simulation and their values are compared. (author/RL)

Robinson's Measure of Agreement as a Parallel Forms Reliability Coefficient

Victor L. Willson

Texas A&M University

Running head: Robinson's measure

Robinson's Measure of Agreement as a Parallel Forms Reliability Coefficient

A major deficiency in classical test theory is the reliance on Pearson product-moment (PPM) correlation concepts in the definition of reliability. PPM measures are totally insensitive to first moment differences in tests which leads to the dubious assumption of essential tan-equivalence. Lord and Novick, (1968; p. 194) suggest that when tests are parallel except for mean difficulty differences the researcher "may prefer some form of the conventional formula (8.8.2)". The formula they present for error variance is

$$\sigma_E^2 = \sigma_Y^2 \left[1 - \rho \left(Y_1, Y_2\right)\right], \qquad (1)$$

estimated by

$$\hat{\sigma}_E^2 = S_y^2 \left(1 - r_{12}\right) \qquad (2)$$

where σ_E^2 = population error variance,

σ_Y^2 = population score variance,

ρ = parallel forms reliability,

S_y^2 = some pooled estimate of S_{y1}^2 and S_{y2}^2

Y_1, Y_2 = random variable score at time 1 or 2

y_1, y_2 = realizations of Y_1, Y_2 at times 1, 2

r_{12} = PPM between y_1, y_2

It is clear that (1) and (2) do not account for nonparallelism in mean difficulty since all parameters and statistics employed are first-moment insensitive. This insensitivity has in recent years been shown to have important consequences. This has been most clearly demonstrated in latent trait models (cf. Hambleton and Cook, [illegible]). Differential parallel test difficulty will affect decisions in criterion referenced testing, mastery testing, and competency testing. Thus, a reliability coefficient that is sensitive to mean difficulty differences is needed.

Procedures

Robinson (1957) proposed a measure of agreement that is sensitive to different test difficulty. He developed it in the contest of K raters but its application to K forms is identical:

$$\rho_a = 1 - \frac{\sum_y \sum_k [Y_{ik} - E(Y_{ik})]^2}{\sum_y \sum_k [Y_{ik} - E \sum_k Y_{ik}]^2} \qquad (3)$$

The sample estimate is

$$\hat{\rho}_a = 1 - \frac{\sum_i \sum_k (y_{ik} - \bar{y}_{i.})^2}{\sum_i \sum_k (y_{ik} - \bar{y}_{..})^2} \qquad (4)$$

where i = ith person

k = kth form, of K forms.

This measure is quite similar to Kelley's (1921) eta-squared statistic except the numerator of (4) is a sum of squares within person across forms pooled across persons. The denominator is the total sum of squares.

Robinson points out that the measure is formally related to the intra class correlation coefficient which both Lord and Novick (1968) and Cronback, Gleser, Nanda, and Rajaratnam (1972) propose in generalizing across subjects (and possibly forms). The relation is as follows (Robinson, 1957):

$$\hat{\rho}_a = \frac{\hat{\rho}_i + 1}{2} \quad \text{for two forms,} \quad (5)$$

$$\hat{\rho}_a = \left(\frac{k-1}{k}\right)\hat{\rho}_i + \frac{1}{k} \quad \text{for k forms.} \quad (6)$$

Computationally $\hat{\rho}_a$ is preferable to the intraclass correlation on a number of grounds: 1) $\hat{\rho}_a$ is always positive or zero, never negative as $\hat{\rho}_i$ may become; 2) it is independent of k, where as $\hat{\rho}_i$ is a function of k; 3) direct tests are available for $\hat{\rho}_a$, since it is a linear function of $\hat{\rho}_i$, for which Fisher (1938) provided distributional tests. Thus, Robinson's measure of agreement complements the generalizability coefficient and gives a practical statistic to estimate reliability in the presence of known form variation in difficulty.

Tests of Significance. From Fisher (1934) the significance test for the intraclass correlation coefficient is given as

$$F = \frac{1 + (n - 1)\hat{\rho}_i}{1 - \hat{\rho}_i} \qquad (7)$$

This F-statistic is compared with a tabled value with k-1 and k (n-1) degrees of freedom for level alpha. This is termed F critical. Then, using (6) and (7), the critical value for $_a$ for significance from zero is

$$\hat{\rho}_a\text{-critical} = \frac{k-1}{k}\left(\frac{F \text{ critical} - 1}{F \text{ critical} + (n-1)}\right) + \frac{1}{k} \qquad (8)$$

Simulation study. A simulation study is presented to acquaint the reader with degree of the coefficient's sensitivity to form difficulty variance. For sets of 50 scores the difficulty of the forms was varied by adding a constant amount to each score in a given form. Results are presented in Tables 1-3 for form internal consistencies of .90, .70, and .50. That is, for internal consistency .90 all forms shared the same two scores which comprised 90% of the within form variance. Each score in the second through sixth form was increased in value 1%, 2%, 5%, or 10% of the total form population variance to produce unequal form means. Robinson's measure of agreement and the intraclass corelation were then computed for each simulation. A total of seventy five runs was made (5 levels of form by 5 levels of mean difference by 3 levels of internal consistency). Inspection of Tables 1 to 3 leads one to conclude that differences are small for highly internally consistent forms (about a .02 difference for coefficient alpha = .90). For forms with moderate internal consistency (.70) the Robinson measure is typically about .05 lower than the intrclass correlation. For low internal consistency (.50) the Robinson measure is typically .12 lower than intraclass correlation for 2 or 3 forms, and it drops to about .07 for 5 or 6 forms. There appears to be no greater difference between the coefficients with greater difference in form means, although the reliability generally drops with greater difference in forms for Robinson's measure. The simulation is merely indicative of the analytical results.

Discussion

Robinson's measure of agreement appears to be a useful alternative to the generalizability coefficient, as it provides a more conservative estimate of reliability under conditions of parallel form differences in mean. This is likely to be especially useful when examining inter rater reliability when internal consistency of the raters is poor. Robinson's measure does not seem advantageous for highly reliable parallel tests such as are encountered in standardized testing programs.

Table 1: Simulation results for Robinson's Measure of Agreement and Intraclass Correlation Coefficient Alpha = .90 for each Form.

Form Differences as % of σ^2	2	3	4	5	6
0%	=.966 =.983	.[illegible]27 .951	.930 .947	.923 .939	.918 .932
1%	.946 .973	.933 .95[illegible]	.927 .945	.930 .944	.905 .921
2%	.949 .975	.92[illegible] .94[illegible]	.910 .932	.924 .939	.925 .937
5%	.960 .980	.9[illegible] .9[illegible]	.92 .94	.899 .919	.912 .927
10%	.971 .986	.8[illegible] .9[illegible]	.908 .931	.916 .933	.837 .864

Note 1: Top number is Robinson's measure of agreement, bottom number is the intraclass correlation for each pair.

Note 2: Each form had 50 observations.

Table 2: Simulation resurlt for Robinson's measure of agreement an intra-class correlation coefficient alpha = .70 for each form.

Form difference as % of σ^2		Number of Forms 2	3	4	5	6
0%	$\hat{\rho}_a$ =	.876	.856	.683	.801	.769
	$\hat{\rho}_i$	.938	.904	.763	.841	.808
1%		.841	.790	.772	.718	.755
		.921	.860	.829	.775	.796
2%		.859	.810	.774	.813	.759
		.929	.873	.830	.850	.799
5%		.872	.810	.717	.764	.788
		.936	.873	.787	.817	.824
10%		.810	.771	.748	.772	.707
		.905	.847	.811	.818	.756

Note 1: Top number is Robinson's measure of agreement, bottom number is the intraclass correlation.

Note 2: Each form had 50 observations.

Table 3: Simulation results for Robinson's measure of agreement and intraclass correlation, coefficient alpha = .50 for each form.

Form difference as % of σ^2		Number of Forms 2	3	4	5	6
0%	$\hat{\rho}_a$ =	.652	.712	.561	.622	.622
	$\hat{\rho}_i$ =	.826	.808	.671	.697	.685
1%		.662	.619	.494	.551	.517
		.831	.746	.620	.641	.597
2%		.829	.605	.630	.633	.606
		.915	.737	.722	.706	.672
5%		.818	.591	.558	.652	.586
		.909	.727	.668	.721	.655
10%		.761	.546	.581	.555	.552
		.881	.697	.686	.644	.626

Note 1: Top number is Robinson's measure of agreement, bottom number is the intraclass correlation.

Note 2: Each form had 50 observations.

REFERENCES

Fisher, R.A. Statistical Methods for Research workers (5th Ed.) Edenburgh: Oliver and Boyd, 1934.

Hambleton, R.K. and Cook, L.L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Lord, F.M. and Novick, M.R. Statistical Theories of Mental Tests. Reading, MA: Addison-Wesley Pub. Co., 1968.

Robinson, W.S. The statistical measure of agreement. American Sociological Review, 1957, 22, 17-25.